WORLD WIDE WEB

WEB PAGES

DOMAIN DATA
10

11 — CRAWLER

16

DATABASE OF
WEB DOMAINS

14

INTERESTING
WEB PAGES
12

48

LOCAL STORAGE
OF INTERESTING
WEB PAGES

FIG. 1

2/3

11

START → DOMAIN TO CRAWL /20 → LOAD HOME PAGE /22 — FAILURE

SUCCESS

IS THIS WEB SITE DUPLICATE OF AN ALREADY VISITED SITE ? /23 — YES

NO /24

IDENTIFY SITE TYPE AND POTENTIAL STRUCTURE /24 ← INITIALIZE "LINKS TO VISIT" TABLE FROM HOME PAGE /26

IS TABLE "LINKS TO VISIT" EMPTY ? /29 — YES — NO → SELECT AND REMOVE LINK FROM "LINKS TO VISIT" TABLE /30 → LOAD WEB PAGE CORRESPONDING TO THIS LINK /27 — FAILURE

SUCCESS

EXAMINE AND CLASSIFY WEB PAGE /28

SITE TIME-OUT OCCURRED ? /25 — NO — YES

COLLECT NEW DOMAINS TO CRAWL /18 ← PRUNE USELESS LINKS; KEEP ONLY INTERESTING LINKS /19 ← COLLECT FROM WEB PAGE NEW INTERNAL LINKS TO VISIT /21

IDENTIFY THE CONTENT OWNER'S NAME FOR THIS SITE /40 → SAVE INTERESTING PAGES FOUND IN THIS SITE /42 → SAVE NEW DOMAINS FOUND IN THIS SITE /44 → END /46

FIG. 2

3/3

58

TABLE: "Links to Visit"

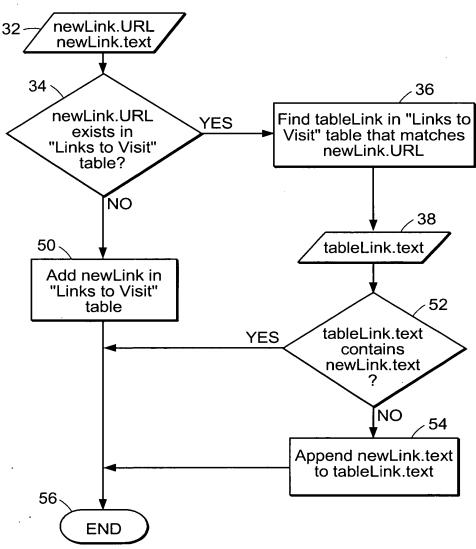| URL	LINK TEXT
http://www.corex.com/index.html	Corex Technologies Corp.
http://www.corex.com/about.html	About Corex
http://www.corex.com/history.html	Company history
http://www.corex.com/products.html	Products > Cardscan Executive
http://www.corex.com/news/nov.html	Press Releases > November 1999
http://www.corex.com/news/dec.html	Press Releases > December 1999

16

32 — newLink.URL
newLink.text

34 — newLink.URL exists in "Links to Visit" table?

YES → 36 — Find tableLink in "Links to Visit" table that matches newLink.URL

NO

50 — Add newLink in "Links to Visit" table

38 — tableLink.text

52 — tableLink.text contains newLink.text ?

YES

NO

54 — Append newLink.text to tableLink.text

56 — END

FIG. 3